RS/6000 SP 375MHz POWER3 SMP High Node

Authors

Bob Amos, Product Manager, Sanjay Deshpande - Senior Engineer,, Mike Mayfield - Senior Technical Staff Member and Frank O'Connell - Senior Technical Staff Member

Abstract

IBM's new 375 MHz POWER3 SMP High Node is the latest addition the RS/6000® SP™ product line. It is the highest performing SP node yet, with double the number of processors, quadruple the maximum memory, double the L2 cache size and significantly higher processor frequency compared with its predecessor, the POWER3 SMP High Node. It also provides an extensive I/O interconnect capability and supports the newest high performance SP Switch2. The new 375 MHz POWER3 SMP High Node's advanced design provides the power, configurability, reliability and growth to support the most demanding computing requirements. It is ideal for deep computing applications such as engineering analysis, molecular modeling or crash simulation or business data analysis such as decision support or enterprise resource planning.

Key attributes of the 375 MHz SMP High Node include:

- Up to sixteen (16) 375 MHz POWER3-II 64-bit processors utilizing IBM's advanced copper-based microprocessor fabrication technology.
- Up to 64GB of high performance memory to support the data requirements of everything from frontier-expanding scientific research to complex decision support environments.
- An industry leading 16GB/s cross-point data switch design providing a fast, cache coherent infrastructure for the interconnection of the processors, memory cards, and I/O components.
- One-cycle latency L1 instruction and data caches per processor to provide immediate access to data and instructions.
- A private 8MB set-associative L2 cache per processor, providing the ability to retain the key data structures of applications and thereby boost application performance significantly.
- Speculative data and instruction prefetching to minimize delays in accessing data and instructions.
- Extensive I/O to support connectivity to storage and networks and the I/O bandwidth to required to handle even the most demanding applications.
- Support for the new high performance SP Switch2
- Reliability, Availability and Serviceability functions and features to support the demands of e-business and long complex engineering and scientific analysis.
- Configuration flexibility which allows processors, memory, I/O and I/O subsystems to be added incrementally as the demand grows.
- Advanced system management and support functions.

This paper provides an overview of how node architecture, system technology, and advanced computer design are brought together to produce outstanding high-performance computational capabilities with high reliability.

What follows is a description of the 375 MHz POWER3 SMP High Node design point, highlighting its advanced performance features. That is followed by a description of the SP I/O Expansion Unit, the POWER3-II microprocessor, and an overview of the advanced reliability features designed into these nodes. Finally, a summary of the system performance is provided.

375 MHz POWER3 SMP High Node Overview

The 375 MHz POWER3 SMP High Node complex consists of a compute node (375 MHz POWER3 SMP High Node) and attached SP Expansion I/O Units. The High Node contains the processors, memory, base I/O and service processor. Up to 6 SP Expansion I/O Units can be optionally attached to the High Node to support I/O demands beyond that provided in the High Node.

The 375 MHz POWER3 SMP High Node is an up to 16-way shared memory multiprocessor, utilizing IBM's advanced POWER3-II copper-based microprocessor. Each microprocessor has a high speed, low latency, 8MB 4-way set-associative L2 cache connected to a dedicated high speed bus running at 250 MHz.

The High Node consists of high bandwidth switched data paths allowing the interconnection of up to 16 POWER3-II processors and of up to 4 I/O elements to two 2-port memory subsystems in a SMP configuration. The processors are organized as nodes (or cards), with 4 processors per card. The I/O elements form a separate node. The processors and I/O elements interface with the system via hardware elements called Node Controllers. The bi-directional data paths between the system elements are point-to-point to allow them to run at a frequency of 125 MHz.

Figure 1 shows the block diagram of the Central Electronics Complex (CEC) of a 375 MHz POWER3 SMP High Node. It is formed of 1 to 4 processor cards, a separate card carrying I/O elements, up to 4 memory cards, and an active back plane which carries the Data Switch and 2 independent memory controllers, each able to support up to 2 of the memory cards. The following table explains the functionality of each of the chips appearing in Figure 1.

Chip	Function
POWER3-II	Copper Based Processor
NCA	Node Controller - Address: Handles address portion of all transactions issued by processors and I/O elements and coordinates data transfers among processors, memory, and I/O elements.
NCD	Node Controller - Data: Contains paths for interprocessor and processor-memory data transfers. Forms a portion of the data switch.
ASX	Address Broadcast
Crossbow	Memory Address Controller. Controls data transfers across paths within the Arrow chips.
Arrow	Memory Data Switch: Carries data to and from memory and between nodes. Forms a portion of the data switch.
Saber	Portal to I/O subsystems.
SP Fabric Adapter	Portal to inter High Node communication interconnect.

375MHz POWER3 SMP High Node
CEC Block Diagram

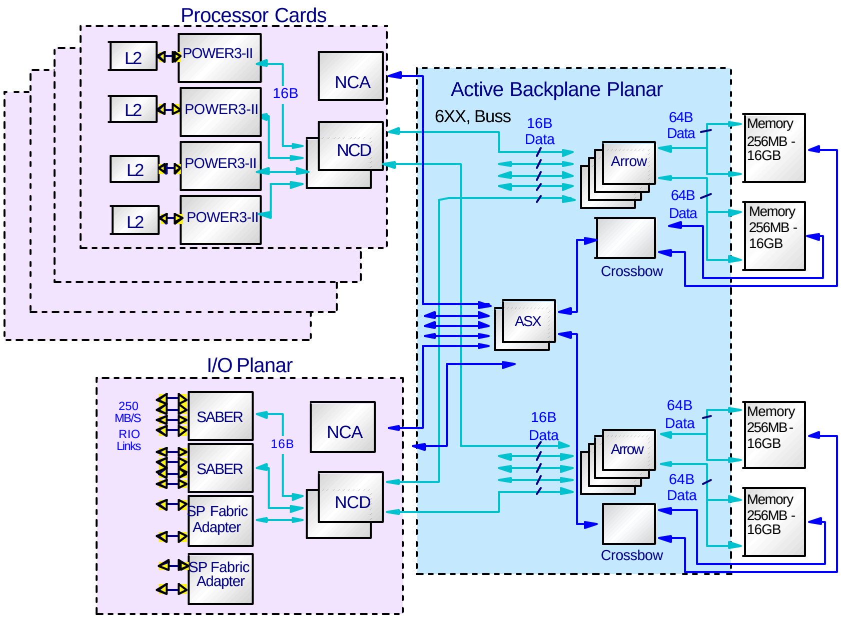


Figure 1.

Each processor card contains four 375 MHz POWER3-II microprocessors. Each processor is supported by 8MB of 4-way set-associative L2 cache on a private high speed bus running at 250 MHz. Four processors are attached to a set of node controller chips through independent point-to-point address and 16 byte wide data buses running at 125 MHz. Each processor can issue up to one address transaction per cycle to the system.

The node controller chips themselves are attached to the rest of the system via a pair of unidirectional address buses and a pair of 16 byte wide bi-directional data buses also running at 125 MHz. This translates to a data bandwidth of 4GB/s per processor card or 16GB/s per system node. The node controller chips provide extensive buffering of addresses and data to improve system performance. Collectively, they also provide coherency management between all of the processors, memory and I/O using a snoopy bus protocol.

Figure 2 shows a block diagram of the physical hierarchy for system memory. Memory is packaged on up to 4 cards connected to the 4 memory ports in the data cross point switch. Memory cards contain either 128MB, 256MB or 512MB SDRAM DIMMs. Each card can contain up to 16GB of memory or up to 64GB for the compute node. 1GB of memory comes in the base system. Memory can be added in 1, 2 or 4GB increments with the addition of cards and pluggable industry standard DIMMs. This provides a highly configurable and upgradable offering, which will grow with the application requirements.

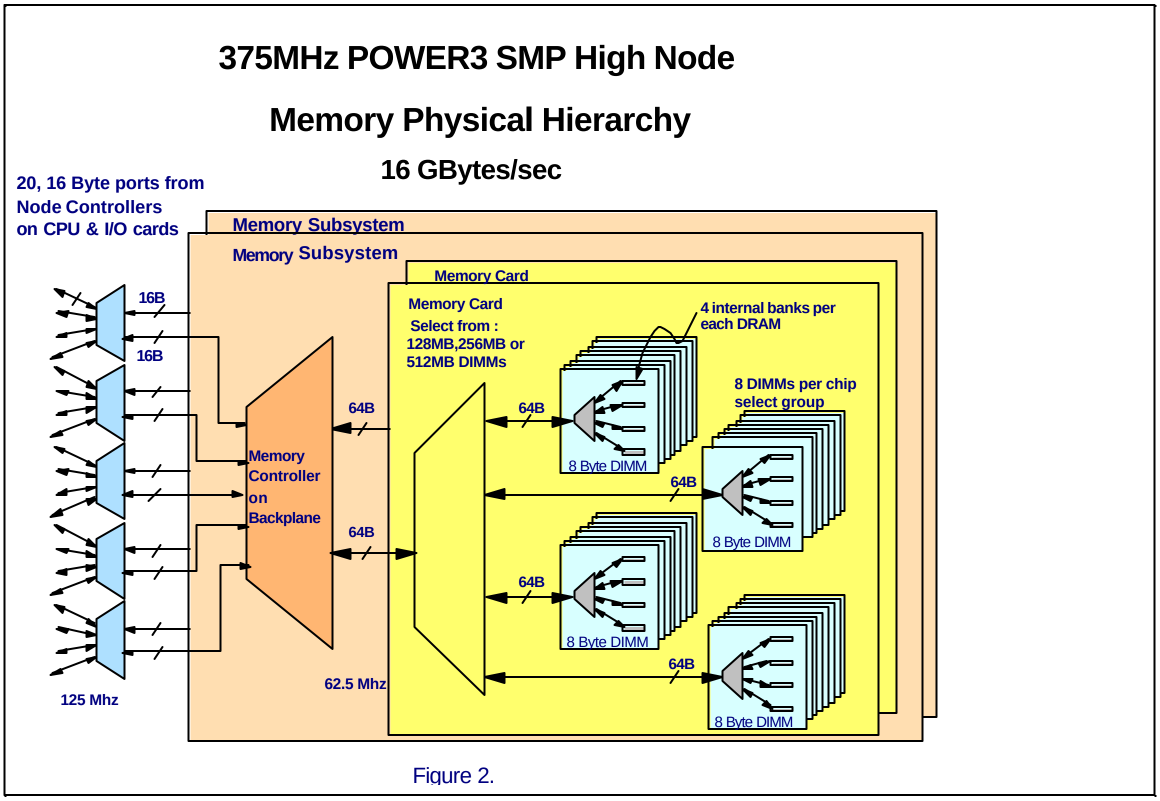


Figure 2.

Each memory subsystem is capable of accepting an address transaction every cycle at 125 MHz. Each memory data port is 64 bytes wide and operates at 62.5 MHz, which translates to a delivered bandwidth of 16GB/s.

Figure 3 shows a block diagram of the I/O subsystem. The figure shows Saber modules, which serve as portals to disk and communication subsystems, and an SP Fabric Adapter which allows a compute node to connect to the SP Switch2 to form a multinode system.

375MHz POWER3 SMP High Node and SP Expansion I/O Unit I/O Topology

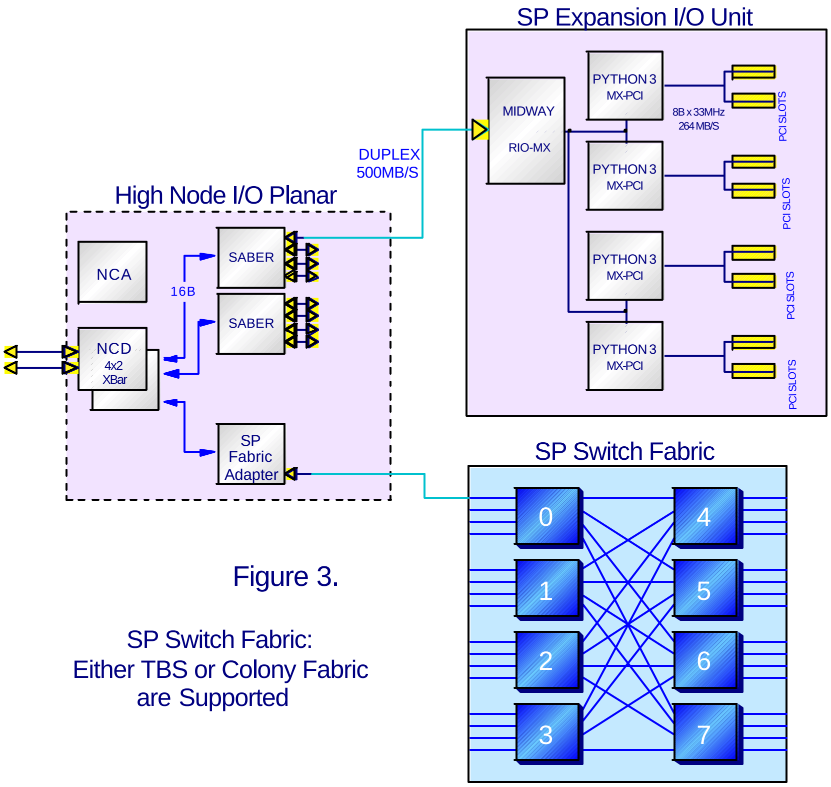


Figure 3.

SP Switch Fabric:
Either TBS or Colony Fabric
are Supported

The High Node I/O consists of four 64bit PCI slots and one 32-bit PCI slot operating at 33 MHz. It provides integrated Ultra SCSI, 10/100 Ethernet , a parallel port, and three serial ports. In addition, there are 6 remote I/O (RIO) connections. These connections allow SP Expansion I/O Units to be connected to the High Node providing incremental growth for applications requiring more I/O connectivity. The RIO ports operate full duplex at 250MB/s each direction (500MB/s total for each link). This provides outstanding I/O bandwidth to each RIO expansion node. With 6 SP Expansion I/O Units connected, a user can obtain a node complex with 53 PCI slots and 26 DASD bays. This combination of High Node and SP Expansion I/O Units into a node complex provides the highest I/O configuration yet on the RS/6000 SP system.

SP Expansion I/O Unit Overview

The SP Expansion I/O Unit is offered in a thin SP node format. It contains eight 64-bit “hot swappable” PCI slots running at 33 MHz and 4 “hot swappable” DASD bays that will accept either SSA or SCSI drives in a variety of sizes.

The node is designed for reliability and easy service. It supports N+1 power and cooling, “hot swappable” mirrored disks drives and PCI slots, redundant links to the 375 MHz POWER3 SMP High Node, and a node supervisor which monitors the health of the system.

Up to 6 expansion units can be attached to the High Node offering a tremendous compute complex with extensive I/O connectivity and performance.

A large variety of PCI adapters are available for the 375 MHZ POWER3 SMP High Node complex, which allow extensive network and storage connectivity and performance. The user may also select from a large variety of IBM’s external storage subsystems, which can be attached to this node complex.

POWER3-II Microprocessor

The POWER3-II microprocessor continues the POWER architecture tradition of bringing real solutions to IBM RS/6000 customers' high-performance compute needs. It supports 64-bit addressability, double-word integer operations, and symmetric multiprocessor configurations.

To satisfy compute intensive requirements, the POWER3-II design contains a highly superscalar core which comprises eight execution units capable of sustaining an execution rate of four instructions per cycle, a 32KB instruction cache, 64KB data cache, and high bandwidth, independent interfaces to the L2 cache and system memory. A block diagram of the POWER3-II processor is show in Figure 4.

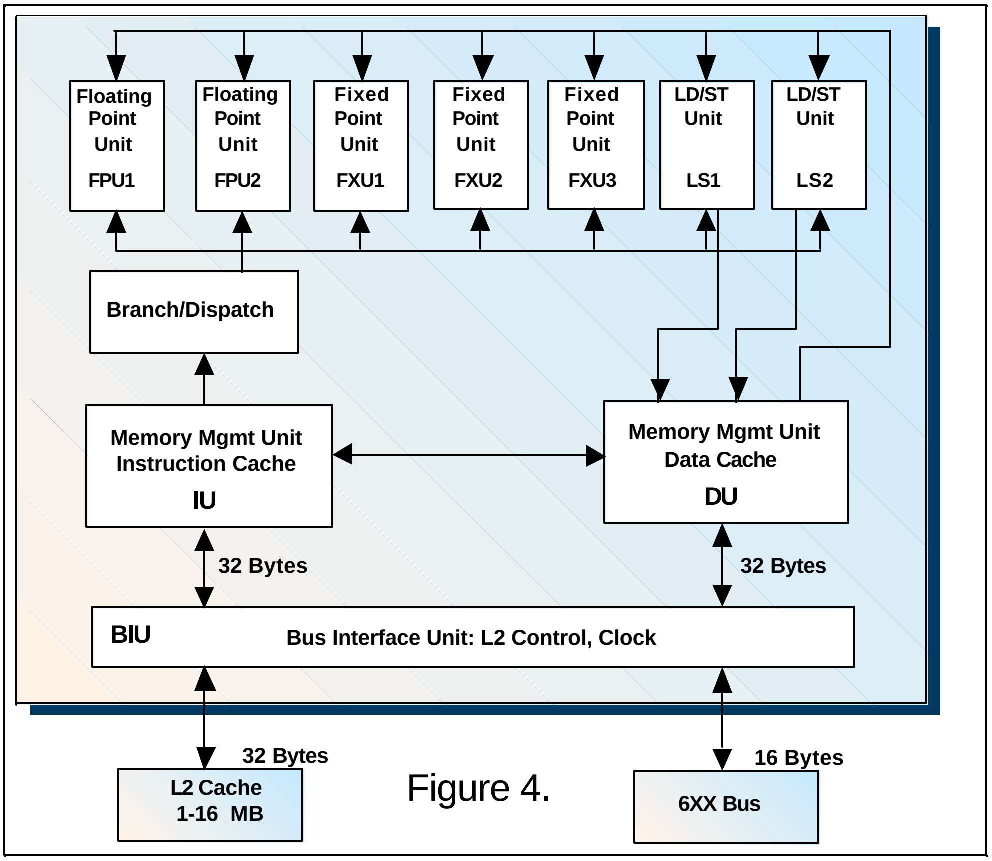


Figure 4.

Two of the three fixed-point execution units (FXUs) provide single-cycle execution for the bulk of the integer arithmetic instructions. The third unit executes the multi-cycle integer instructions such as multiply and divide.

The two floating-point execution units (FPUs) are fully independent, each containing dedicated hardware for square root and divide routines as well as fused multiply-add instruction execution. The FPUs are fully pipelined with three-cycle latency and single-cycle throughput.

Two load/store units provide the data movement capability to support the three fixed point and two floating point execution units. A 16-entry store queue buffer prevents stores from stalling the machine while loads are being performed. Loads are also executed speculatively, improving data throughput.

The branch execution unit employs dynamic branch prediction, with four pending predicted branches supported. The branch target address cache contains 256 entries (128 by 2-way associative), and the branch history table has 2048 entries.

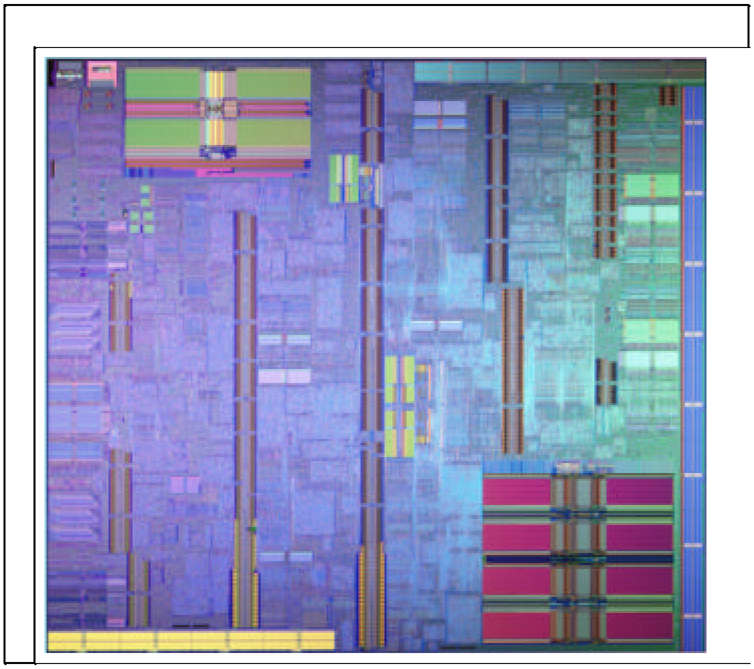
The instructions are speculatively executed with a unique register renaming scheme that involves a total of 64 virtual rename registers (32 fixed and 32 floating-point), and a total of 40 physical rename registers (16 fixed point and 24 floating-point).

The on-board BIU contains the interface logic supporting up to 16Mbytes L2, 6XX system bus protocols, and dedicated hardware to reduce latency to memory.

Containing 25 million transistors, the POWER3-II processor die is shown in Figure 5. It is manufactured in IBM's 0.22 micron hybrid CMOS 7S technology, with five levels of copper interconnect metallurgy.

To ensure that potentially needed data and instructions are available to keep the core from stalling, the POWER3-II processor designers invested in two key latency reduction techniques:First, all caches are non-blocking. The instruction cache supports two outstanding misses, and the data cache supports up to four outstanding misses

Second, the POWER3-II processor implements sequential instruction and data access detection algorithms in hardware, which permit the prefetching of cache lines to closer levels of the memory hierarchy. The POWER3-II processor prefetches up to four separate data streams with a depth of two lines for each stream. Prefetching data significantly reduces the apparent memory latency and improves the data bandwidth, thus increasing performance and processor utilization. Instructions are prefetched into the L1 cache up to one sequential line ahead of the line currently being accessed on the predicted path, which often eliminates delays with instruction fetches from memory.



Reliability, Availability and Serviceability Features

The 375 MHz POWER3 SMP High Node and SP Expansion I/O Units were designed with high RAS capabilities in mind. From memory to I/O to processors this node represents significant advancements in system RAS.

Memory and L2 cache are protected with single bit correct, double bit detect ECC. L1 cache is parity protected. Memory incorporates scrubbing and supports continued operation with a full memory chip failure as further protection.

Both the 375 MHz POWER3 SMP High Nodes and the SP Expansion I/O Unit incorporate N+1 power and N+1 cooling. The Expansion Units also provide redundant links to the High Node.

Disk mirroring is standard on both the High Node and the Expansion Unit. It provides redundant storage allowing continued operation in the presence of a disk failure.

The Expansion Unit provides “hot swappable” capability for both SCSI and SSA disks and PCI adapters, which often enables maintenance concurrent with node operation. This reduces down time for maintenance.

CPU and Memory RepeatGuard capability is provided. This function checks for excessive soft or hard fails at boot time and deconfigures a faulty memory bank or processor for deferred repairs. CPU Guard also provides for dynamic CPU deallocation when excessive soft fails are detected. The soft fail threshold is user selectable.

Performance

The 375 MHz POWER3 SMP High Node's superior computing performance is the result of clearly articulated design objectives shaped by the characteristics of challenging customer applications, and knowledge from years of experience in designing the RS/6000 family of POWER and POWER2 processors. Consistent with its POWER and POWER2 heritage, the 375MHz POWER3 SMP High Node distinguishes itself from its competitors in its ability to deliver main memory bandwidth, a crucial attribute to many applications on the high performance computing frontier. These applications can achieve a quantum leap in performance on the High Node and, in turn, provide new scientific insights and competitive advantages to their users. This is most clearly demonstrated in the ASCI White Supercomputer, which posted a result of over 4.9 Teraflops on the LINPACK HPC benchmark utilizing 464 375 MHz POWER3 SMP High Nodes.

The 375 MHz POWER3 SMP High Node excels in data delivery performance; indeed this was its most significant and most challenging design objective. The most sophisticated features of its design are all integral pieces of its data delivery system: dual load/store execution units, an interleaved data cache, multiple-outstanding cache-miss support, wide data paths to memory and its L2 cache, hardware data prefetch, node controllers, and a non-blocking data cross point switch.

Figure 6 shows the history of POWER, POWER2, and POWER3 STREAM performance to date, including the 375 MHz POWER3 SMP High Node, in the STREAM benchmark, a standard measurement of sustained memory bandwidth. Note that the 375 MHz POWER3 SMP High Node provides a significant jump over the already impressive performance of the POWER3 SMP High Node and POWER3 SMP Thin and Wide Nodes. As in the past, this new RS/6000 system once again sets the industry standard for sustained memory bandwidth in systems of its class.

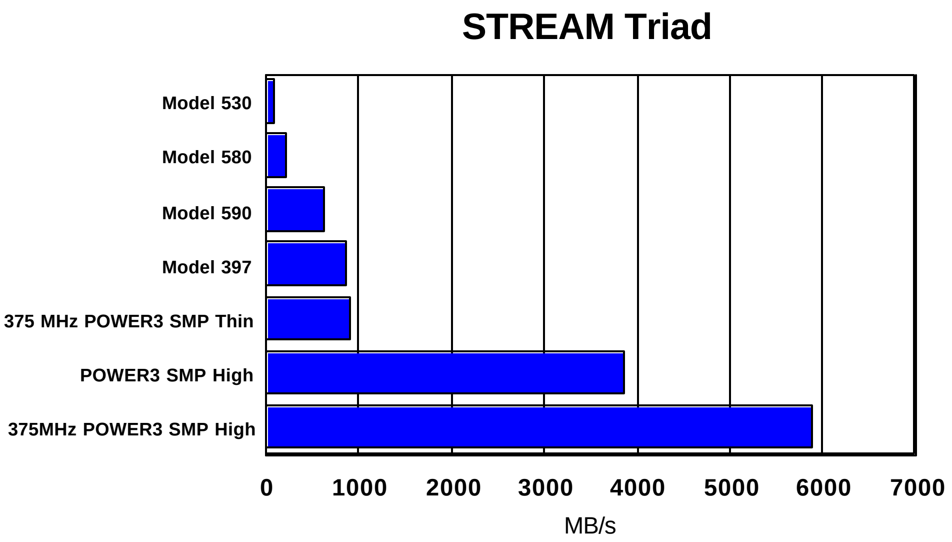


Figure 6

Complimenting the 375 MHz POWER3 SMP High Node's excellent memory bandwidth are floating-point and fixed-point execution units that can sustain remarkable rates of computation. Figure 7 shows the LINPACK 1000 performance for the 375 MHz POWER3 High Node and a sample of its product pedigree. This benchmark measures the average performance to factorize and solve a dense matrix of rank 1000 and is representative of large amounts of work found in many important applications. Just as with sustained bandwidth, the High Node greatly improves upon a distinguished record of floating-point performance.

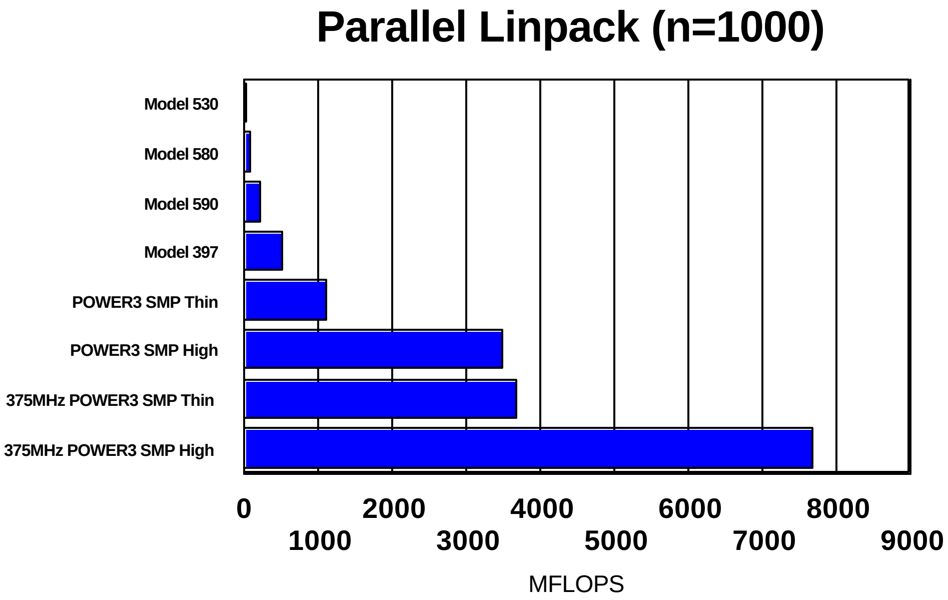


Figure 7

The 375 MHz POWER3 High Node also demonstrates its potential in transaction-oriented computing environments with the ROLTP metric in Figure 8, which estimates its OLTP performance in a relative sense. Its 16-way high performance cache-memory design is well balanced to excel in any combination of instruction and/or data intensive environments. This versatility makes the High Node the perfect system for enterprise-wide commercial computing and e-business environments.

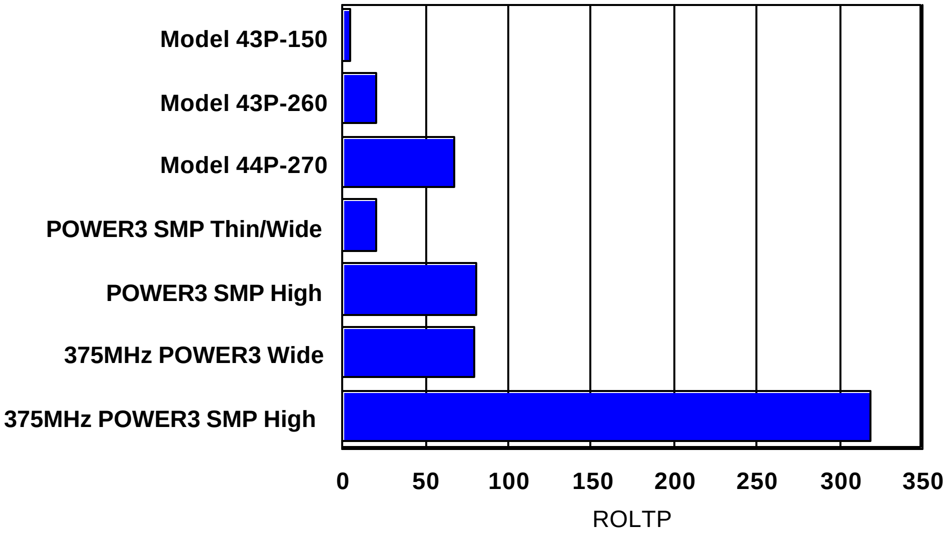


Figure 8

Summary

The 375 MHz POWER3 SMP High Node, along with the SP Expansion I/O Unit, represents a powerful and exciting new offering for the RS/6000 SP product line. It significantly extends the compute capability over its predecessor node by utilizing 16 of the new, high performance copper-based Power 3-II microprocessors, quadrupling the maximum memory, adding a 4-way set-associative 8MB L2 cache, and adding support for the new SP Switch2. And finally, the upgrade to the 375 MHz High Node from the POWER3 SMP High Node is as easy as swapping the processor cards.

Biographies

Bob Amos is a Product Manager. Sanjay Deshpande is a Senior Engineer. Mike Mayfield is a Senior Technical Staff Member. Frank O'Connell is a Senior Technical Staff Member. All authors are members of IBM Server Group in Austin, Texas.

Special Notices

This document was produced in the United States. IBM may not offer the products, programs, services or features discussed herein in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the products, programs, services, and features available in your area. Any reference to an IBM product, program, service or feature is not intended to state or imply that only IBM's product, program, service or feature may be used. Any functionally equivalent product, program, service or feature that does not infringe on any of IBM's intellectual property rights may be used instead of the IBM product, program, service or feature.

Information in this document concerning non-IBM products was obtained from the suppliers of these products, published announcement material or other publicly available sources. Sources for non-IBM list prices and performance numbers are taken from publicly available information including D.H. Brown, vendor announcements, vendor WWW Home Pages, SPEC Home Page, GPC (Graphics Processing Council) Home Page and TPC (Transaction Processing Performance Council) Home Page. IBM has not tested these products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local IBM office or IBM authorized reseller for the full text of a specific Statement of General Direction.

The information contained in this document has not been submitted to any formal IBM test and is distributed "AS IS". While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. The use of this information or the implementation of any techniques described herein is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. Customers attempting to adapt these techniques to their own environments do so at their own risk.

IBM is not responsible for printing errors in this publication that result in pricing or information inaccuracies.

The information contained in this document represents the current views of IBM on the issues discussed as of the date of publication. IBM cannot guarantee the accuracy of any information presented after the date of publication.

All prices shown are IBM's suggested list prices; dealer prices may vary.

IBM products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Information provided in this document and information contained on IBM's past and present Year 2000 Internet Web site pages regarding products and services offered by IBM and its subsidiaries are "Year 2000 Readiness Disclosures" under the Year 2000 Information and Readiness Disclosure Act of 1998, a U.S statute enacted on October 19, 1998. IBM's Year 2000 Internet Web site pages have been and will continue to be our primary mechanism for communicating year 2000 information. Please see the "legal" icon on IBM's Year 2000 Web site (www.ibm.com/year2000) for further information regarding this statute and its applicability to IBM.

Any performance data contained in this document was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

The following terms are registered trademarks of International Business Machines Corporation in the United States and/or other countries: ADSTAR, AIX, AIX/6000, AS/400, C Set++, CICS, CICS/6000, DB2, ESCON, IBM, Information Warehouse, LANStreamer, LoadLeveler, Magstar, MediaStreamer, Micro Channel, MQSeries, Netfinity, Parallel Sysplex, POWERparallel, PowerPC (logo), RS/6000, S/390, Service Director, ThinkPad, TURBOWAYS, Videocharger, VisualAge. The following terms are trademarks of International Business Machines Corporation in the United States and/or other countries: AIX PVMe, AS/400e, DB2 Universal Database, Deep Blue, e-business (logo), HACMP/6000, Intelligent Miner, Intellistation, Network Station, POWER2 Architecture, PowerPC 604, PowerPC Architecture, SmoothStart, SP. A full list of U.S. trademarks owned by IBM may be found at www.ibm.com/legal/copy/trade.html.

Notes on Benchmarks and Values

The benchmarks and values shown here were derived using particular, well configured, development-level computer systems. Unless otherwise indicated for a system, the values were derived using 32-bit applications and external cache, if external cache is supported on the system. All benchmark values are provided "AS IS" and no warranties or guarantees are expressed or implied by IBM. Actual system performance may vary and is dependent upon many factors including system hardware configuration and software design and configuration. Buyers should consult other sources of information to evaluate the performance of systems they are considering buying and should consider conducting application oriented testing. For additional information about the benchmarks, values and systems tested, contact your local IBM office or IBM authorized reseller or access the following on the Web:

TPC	http://www.tpc.org
GPC	http://www.spec.org/gpc
SPEC	http://www.spec.org
Pro/E	http://www.proe.com
Linpack	http://www.netlib.no/netlib/benchmark/performance.ps
Notesbench Mail	http://www.notesbench.org

Unless otherwise indicated for a system, the performance benchmarks were conducted using AIX 4.3, and AIX XL FORTRAN V6.1. with optimization where the compilers were used in the benchmark tests.

The Linpack benchmark reflects the performance of the microprocessor, memory architecture, and compiler of the tested system.
- LINPACK TPP (Toward Peak Performance) - n=1,000 is the array size. The results are measured in MFLOPS.
Linpak Benchmarks from: http://performance.netlib.org/performance/html/PDSreports.html**.**

STREAM is a program which J. McCalpin of University of Virginia developed and measures sustainable memory bandwidth (in MB/s) and the corresponding computation rate for simple vector kernels. The results reported in this paper are the fastest TRIAD program using a uniprocessor machine. STREAM Benchmark from: http://www.cs.virginia.edu/stream/.